Suboptimal Behavior of Bayes and MDL in Classification under Misspecification

Peter Grünwald and John Langford

Optimality of Bayes

Prior = distribution over data generating distributions

If: Distribution generating data drawn from prior

Then: Use Bayes law to find posterior

$$\Pr(\text{Distribution}|\text{Data}) = \frac{\Pr(\text{Data}|\text{Distribution})\text{Prior}(\text{Distribution})}{\Pr(\text{Data})}$$

For classification, predict according to:

$$c(x) = I\left(\sum_{\text{Distribution}} \Pr(\text{Distribution}|\text{Data}) \Pr(y=1|x, \text{Distribution}) > 0.5\right)$$

Basic Bayes Background

Suppose data drawn i.i.d. from D.

1. If Prior(D)>0 and <mild conditions> then posterior converges to D. (Doob, Blackwell, and Dubins)

2. If Prior(D)=0 and <stronger conditions> then posterior converges to D': Prior(D')>0 and KL($D'||D$) minimized. (Schwartz and others)

This result: <stronger conditions> do not hold for classification and (2) does not hold.

Definition of Classification Problem

X = input space

$Y = \{0, 1\}$ = prediction problem

D = distribution on $X \times Y$

$c : X \to Y$ = classifier

$e_D(c) = \mathsf{Pr}_{x,y\sim D}(c(x) \neq y)$ = classifier error rate

S = m i.i.d samples from D

$\hat{e}_S(c) = \sum_{x,y\sim S} I(c(x) \neq y)$ = number of errors on sample S

Goal: Minimize $e_D(c)$ given i.i.d. samples from D

Basic Result

Theorem: There exists X, D, "Prior" on classifiers,

$$\lim_{m\to\infty} \Pr_{S\sim D^m}\left(e_D(c_{\mathsf{Bayes}}) > 0.1 + \inf_{c:\mathsf{Prior(c)}>0} e_D(c)\right) = 1$$

$$\lim_{m\to\infty} \Pr_{S\sim D^m}\left(e_D(c_{\mathsf{MDL}}) > 0.1 + \inf_{c:\mathsf{Prior(c)}>0} e_D(c)\right) = 1$$

($c_{\mathsf{Bayes}}, c_{\mathsf{MDL}}$ = classifiers learned by algorithms explained next)

Contrasting Result

Theorem: For all X, D, "Prior" on classifiers, $K > 0$

$$\lim_{m\to\infty} \Pr_{S\sim D^m}\left(e_D(c_{\mathrm{Occam}}) > K + \inf_{c:\mathrm{Prior(c)}>0} e_D(c)\right) = 0$$

where

$$c_{\mathrm{Occam}} = \arg\min_{c:P(c)>0} \hat{e}_S(c) + \sqrt{\frac{\ln\frac{1}{P(c)} + \ln\frac{1}{\delta}}{2m}}$$

Outline

1. Define exact algorithms

2. Present Proof (It's easy!)

3. Discuss

Defining c_{MDL} aka "two part code MDL"

$$c_{\mathsf{MDL}} = \arg\min_{c:P(c)>0} \log \frac{1}{P(c)} + \log \binom{m}{\hat{e}_S(c)}$$

= classifier with minimal description length + error description length

Defining c_{Bayes}

Let $P(y^m|c,\vartheta,x^m) = \vartheta^{\hat{e}_S(c)}(1-\vartheta)^{m-\hat{e}_S(c)}$

= probability of error sequence given noise rate ϑ.

$P(\vartheta) = \text{Uniform}(0,1)$

$P(c,\vartheta|y^m,x^m) = \frac{P(y^m|c,\vartheta,x^m)P(c)P(\vartheta)}{P(y^m|x^m)}$ = posterior

$$c_{\text{Bayes}}(x) = \text{Sign}\left(\sum_{c,\vartheta} P(c,\vartheta|y^m,x^m)\vartheta c(x)\right)$$

= weighted majority according to posterior.

c_{Bayes} based on logistic regression: $\frac{1}{1+e^{\beta c(x)}}$

- Standard means to convert real number into "probability"
- Equivalent to "label noise", $\Pr_{(x,y)\sim D}(y \neq c(x))$ (independent of x)
- Used for Bayesian version of neural nets and SVMs (Tipping 2002)
- Known as "not the right thing to do", but difficult to find alternative when noise process unknown.

Proof is Constructive

Let $X = \{0, 1\}^*$, $Y = \{0, 1\}$

Let $c_i(x) = x_i$

To draw from D:

1. pick y uniformly

2. pick x_i s.t. $\Pr(x_i \neq y) = 0.3$ for $i \geq 1$.

3. pick x_0 s.t. $\Pr(x_0 \neq y) = 0.2$.

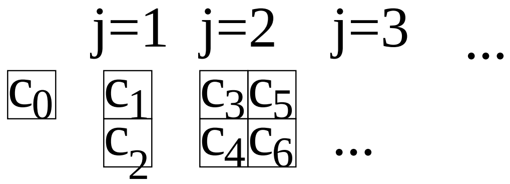
(All randomization is independent.)

Proof, the choice of prior

$$P(c_i) = \frac{1}{2^{\log i + O(\log \log i)}}$$

Method:

1. Divide classifiers into blocks of size 2^j.



2. for i in codebook j, $P(c_i) = \frac{1}{2j(j+1)2^j}$ and $P(c_0) = \frac{1}{2}$

Proof, working through the numbers

code length of $c_0 \geq \log \binom{m}{0.2m} \simeq mH(0.2)$

code length of first classifier with 0 error $\simeq m \log \frac{1}{0.7}$

first classifier with 0 error always preferred.

Pick $|S| = m = 100$, then:

$mH(0.2) = 72.19$

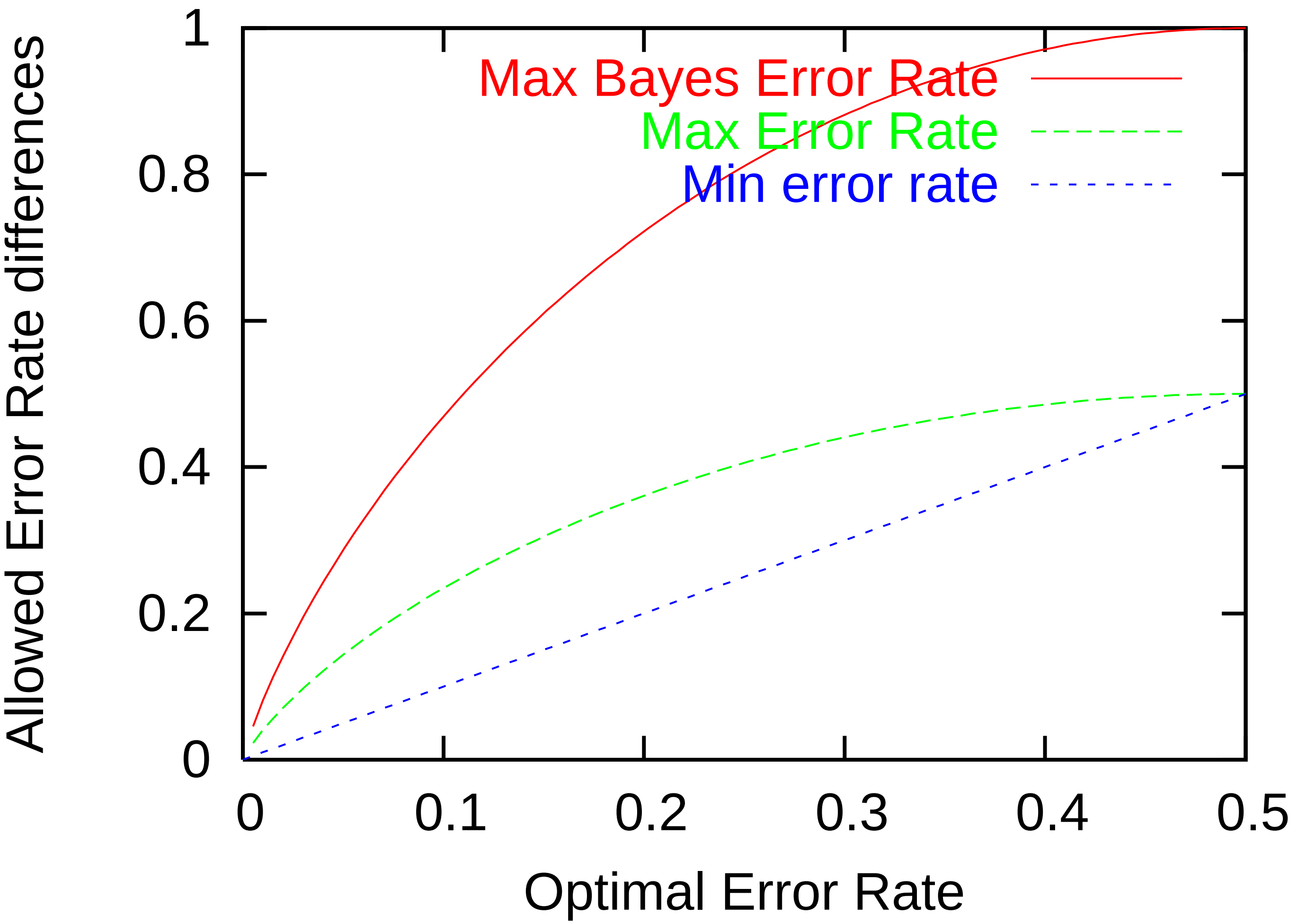
$m \log \frac{1}{0.7} = 51.45$

Proof, other counterexamples

For c_{Bayes}, need to change example. Let $\frac{1}{2}$ of examples be "hard" (error rate 0.6) and the other half be "easy" (error rate 0).

$\Rightarrow$ majority vote on "hard" examples converges to always predicting wrong.

- The proof also applies to KL-divergence!
- Applies to many variants of these techniques
- Applies in a Objective sense: In randomization with respect to prior over classifiers.



Max Bayes Error Rate
Max Error Rate
Min error rate
Allowed Error Rate differences
Optimal Error Rate
0
0.2
0.4
0.6
0.8
1
0
0.1
0.2
0.3
0.4
0.5

One Interpretation: Comparing two methods of coping with lack of knowledge